A Theoretical Understanding of Gradient Bias in Meta-Reinforcement Learning

Bo Liu[1,*], Xidong Feng[2,*],
Jie Ren[3], Luo Mai[3], Rui Zhu[4],
Haifeng Zhang[1], Jun Wang[2],
Yaodong Yang[5],
[1]Institute of Automation, CAS, [2]University College London,
[3]University of Edinburgh, [4]DeepMind,
[5]Peking University

Problem Setting

- GMRL
  - Gradient-based Meta-Reinforcement Learning

$$\max_{\boldsymbol{\phi}} J^K(\boldsymbol{\phi}) := J^{\mathrm{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K),$$

$$\text{s.t. } \boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i + \alpha \nabla_{\boldsymbol{\theta}^i} J^{\mathrm{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i), i \in \{0, 1 \ldots K-1\}$$

$\boldsymbol{\theta}$ are inner-loop policy parameters, $\boldsymbol{\phi}$ are meta parameters,
α is the learning rate,
J^{In} and J^{Out} are value functions for the inner and the outer-loop learner

Problem Setting

- GMRL
  - Gradient-based Meta-Reinforcement Learning

Table 1: Four typical gradient-based Meta-RL (GMRL) algorithms.

Category	Algorithms	Meta parameter ϕ	Inner parameter θ
Few-shot RL	MAML [10]	Initial Parameter	Initial Parameter
Opponent Shaping	LOLA [13]	Ego-agent Policy	Other-agent Policy
Single-lifetime MGRL	MGRL [39]	Discount Factor	RL Agent Policy
Multi-lifetime MGRL	LPG [26]	LSTM Network	RL Agent Policy

Problem Setting

- Meta-gradient Estimation
  - Proposition 3.1 (K-step Meta-Gradient).

$$\nabla_{\boldsymbol{\phi}} J^K(\boldsymbol{\phi}) = \nabla_{\boldsymbol{\phi}} J^{Out}(\boldsymbol{\phi}, \boldsymbol{\theta}^K) + \alpha \nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}^K \nabla_{\boldsymbol{\theta}^K} J^{Out}(\boldsymbol{\phi}, \boldsymbol{\theta}^K)$$

  - $\nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}^K$ takes the form:

$$\nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}^K = \sum_{i=0}^{K-1} \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{In}(\boldsymbol{\phi}, \boldsymbol{\theta}^i) \prod_{j=i+1}^{K-1} \left(I + \alpha \nabla^2_{\boldsymbol{\theta}^j} J^{In}(\boldsymbol{\phi}, \boldsymbol{\theta}^j) \right)$$

Motivation

- Existing Meta-gradient Estimation is **biased**:
  - Compositional Bias in: $\nabla_\phi \hat{J}^{\text{Out}}(\phi, \hat{\theta}^K, \tau_3)$, $\nabla_{\hat{\theta}^K} \hat{J}^{\text{Out}}(\phi, \hat{\theta}^K, \tau_3)$
  - Multi-step Hessian Bias in: $\nabla^2_{\hat{\theta}^j} J^{\text{In}}(\phi, \hat{\theta}^j, \tau_2^j)$

Analytical Form of K-step Meta-Gradient Estimate:

$$\nabla_\phi \hat{J}^K(\phi) = \nabla_\phi \hat{J}^{\text{Out}}(\phi, \hat{\theta}^K, \tau_3) + \alpha \nabla_\phi \hat{\theta}^K \nabla_{\hat{\theta}^K} \hat{J}^{\text{Out}}(\phi, \hat{\theta}^K, \tau_3)$$

$\nabla_\phi \hat{\theta}^K$ takes the form:

$$\nabla_\phi \hat{\theta}^K = \sum_{i=0}^{K-1} \nabla_\phi \nabla_{\hat{\theta}^i} J^{\text{In}}(\phi, \hat{\theta}^i, \tau_1^i) \prod_{j=t+1}^{K-1} \left(I + \alpha \nabla^2_{\hat{\theta}^j} J^{\text{In}}(\phi, \hat{\theta}^j, \tau_2^j)\right)$$

Analysis

- **Compositional Bias**
  - Consider a non-linear compositional scalar objective $f(\boldsymbol{\theta}^K)$, the gradient estimation bias comes from the fact that：

$$f(\boldsymbol{\theta}^K) = f(\mathbb{E}[\hat{\boldsymbol{\theta}}^K]) \neq \mathbb{E}[f(\hat{\boldsymbol{\theta}}^K)]$$

  - If one substitutes the non-linear function $f(\boldsymbol{\theta}^K)$ with $J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K)$, then a typical meta-gradient estimation in GMRL introduces compositional bias:

$$\mathbb{E}[\nabla_{\hat{\boldsymbol{\theta}}^K} \hat{J}^{\text{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \tau_3)] = \mathbb{E}[\nabla_{\hat{\boldsymbol{\theta}}^K} J^{\text{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K)] \neq \nabla_{\boldsymbol{\theta}^K} J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K)$$

$$\mathbb{E}[\nabla_{\boldsymbol{\phi}} \hat{J}^{\text{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \tau_3)] = \mathbb{E}[\nabla_{\boldsymbol{\phi}} J^{\text{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K)] \neq \nabla_{\boldsymbol{\phi}} J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K)$$

Analysis

- **Compositional Bias**

Lemma 4.4 (Compositional Bias). *Suppose that Assumption 4.1 and 4.2 hold, let* $\hat{\Delta}_C = \mathbb{E}[\|f(\hat{\boldsymbol{\theta}}^K) - f(\boldsymbol{\theta}^K)\|]$ *be the compositional bias and* C_0 *the Lipschitz constant of* $f(\cdot)$, $|\boldsymbol{\tau}|$ *denote number of trajectories used to estimate inner-loop gradient in each inner-loop update step,* α *the learning rate, then we have,*

$$\hat{\Delta}_C \leq C_0 \mathbb{E}[\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\|] \leq C_0 \left((1 + \alpha c_2)^K - 1\right) \frac{\hat{\sigma}_{In}}{c_2 \sqrt{|\boldsymbol{\tau}|}}, \tag{6}$$

where $\hat{\sigma}_{In} = \max_i \sqrt{\mathbb{V}[\nabla_{\theta^i} \hat{J}^{In}(\boldsymbol{\phi}, \boldsymbol{\theta}^i, \boldsymbol{\tau}_0^i)]}$, $i \in \{0, .., K-1\}$.

- Lemma 4.4 indicates that the compositional bias comes from the inner-loop policy gradient estimate, concerning learning rate α, sample size $|\boldsymbol{\tau}|$ and variance of policy gradient estimator $\hat{\sigma}_{In}$.

- This is a fundamental issue in many existing GMRL algorithms since applying stochastic policy gradient update can introduce estimation error.

Analysis

- **Multi-step** Hessian Bias
  - For one-step Meta-Gradient, $\nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}^K$ takes the form:

$$\nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}^1 = \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^1} J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^1)$$

  - For K-step Meta-Gradient, $\nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}^K$ takes the form:

$$\nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}^K = \sum_{i=0}^{K-1} \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\text{In}}\left(\boldsymbol{\phi}, \boldsymbol{\theta}^i\right) \prod_{j=t+1}^{K-1} \left(I + \alpha \nabla^2_{\boldsymbol{\theta}^j} J^{\text{In}}\left(\boldsymbol{\phi}, \boldsymbol{\theta}^j\right)\right)$$

Analysis

- **Multi-step** Hessian Bias

Theorem 4.5 (Upper bound for the bias). *Suppose that Assumption 4.1 and 4.2 and 4.3 hold. Let $J_{\phi,\theta}$ denote $\nabla_\phi \nabla_\theta J^{In}$, $H_{\theta,\theta}$ denote $\nabla_\theta^2 J^{In}$, $\hat{\Delta}^K = \|\mathbb{E}[\nabla_\phi \hat{J}^K(\phi)] - \nabla_\phi J^K(\phi)\|$ be the meta-gradient estimation bias, set $B = 1 + \alpha c_2$. Then the bound of bias hold:*

$$\hat{\Delta}^K \leq \mathcal{O}\left((B + \hat{\Delta}_H)^{K-1}\left(\mathbb{E}[\|\hat{\theta}^K - \theta^K\|] + \hat{\Delta}_J + (K-1)\right)\right). \quad (9)$$

  - The multi-step Hessian bias has polynomial impact on upper bound of meta-gradient bias.

Understanding Existing Mitigations

- Mitigation for Compositional Bias
  - Compositional bias caused by the estimation error of inner-loop policy gradient.
  - We propose to use off-policy learning technique to handle the compositional bias problem by reusing samples.

$$\mathbb{E}_{\boldsymbol{\tau} \sim p(\boldsymbol{\tau};\theta)} \left[\sum_{t=0}^{H-1} \frac{\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\mu(\boldsymbol{a}_t|\boldsymbol{s}_t)} \mathcal{R}_{\boldsymbol{\phi}}(\boldsymbol{\tau}) \right]$$

Understanding Existing Mitigations

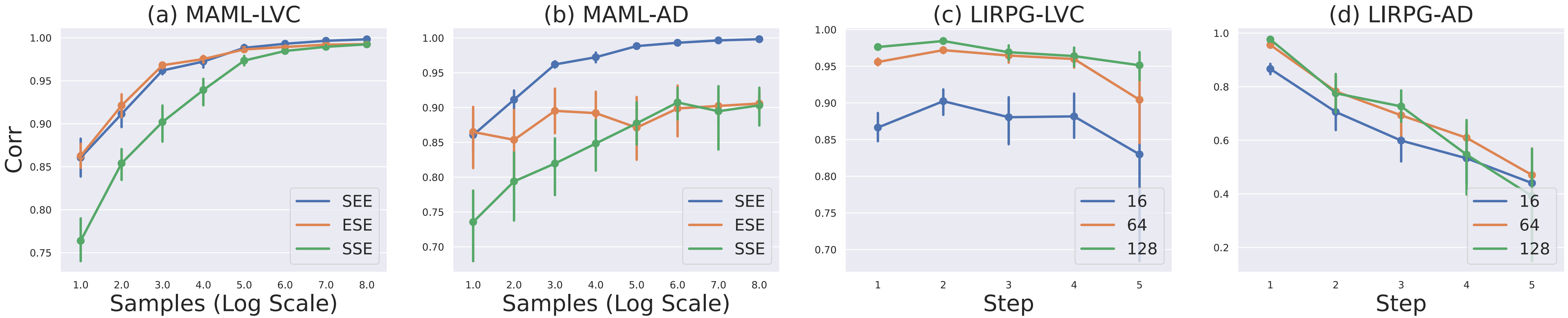
- Mitigation for Multi-step Hessian Bias
  - Hessian estimation bias can **significantly** increase meta-gradient estimation bias in multi-step inner-loop setting.
  - We apply the Low Variance Curvature (LVC) method (Rothfuss et al., 2018) beyond the scope of MAML-RL.

$$\nabla_\theta^2 J_{\mathrm{LVC}}^{\mathrm{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \nabla_\theta \mathbb{E}_{\boldsymbol{\tau} \sim p(\boldsymbol{\tau};\boldsymbol{\theta})} \left[\sum_{t=0}^{H-1} \frac{\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t | \boldsymbol{s}_t)}{\perp \pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t | \boldsymbol{s}_t)} \mathcal{R}_{\boldsymbol{\phi}}(\boldsymbol{\tau}) \right]$$

  - where ⊥ is the stop-gradient operation and it detaches the gradient dependency from the computation graph.

Empirical Results

- Tabular MDP

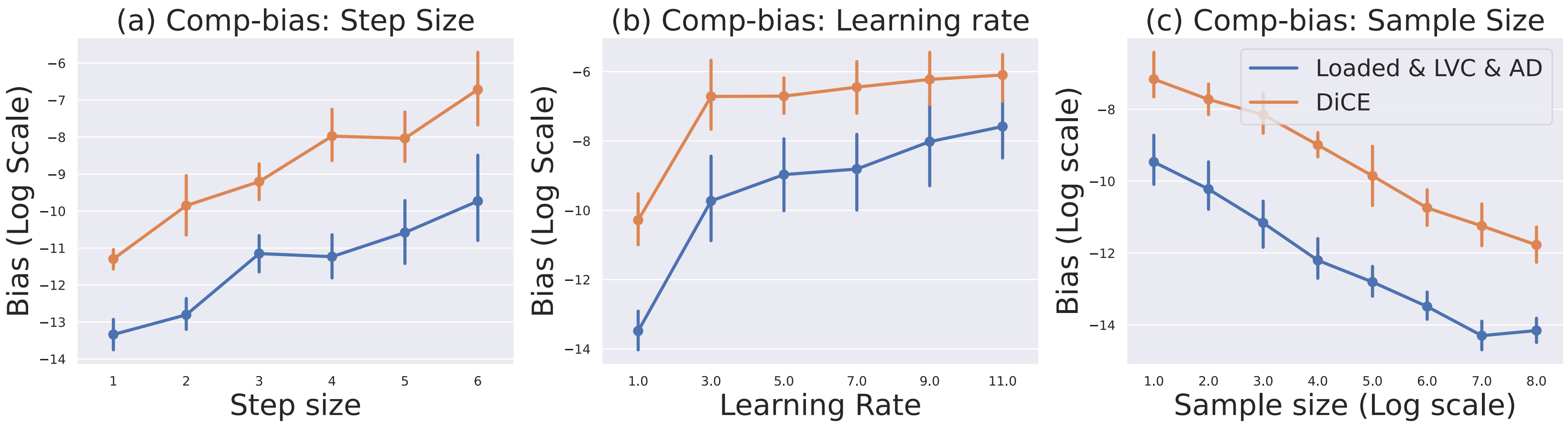


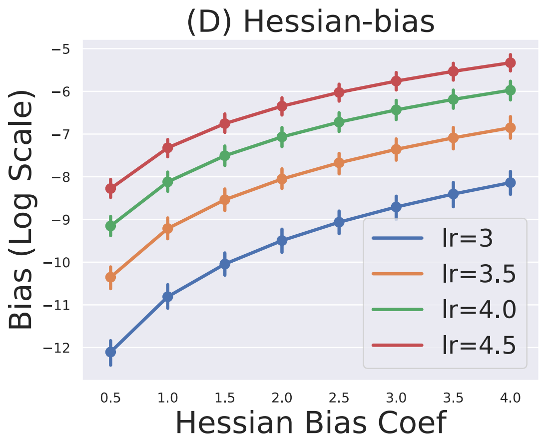
Ablation study on sample size and estimators in MAML-RL

Ablation study on sample size, steps and estimators in LIRPG

Empirical Results

- Tabular MDP



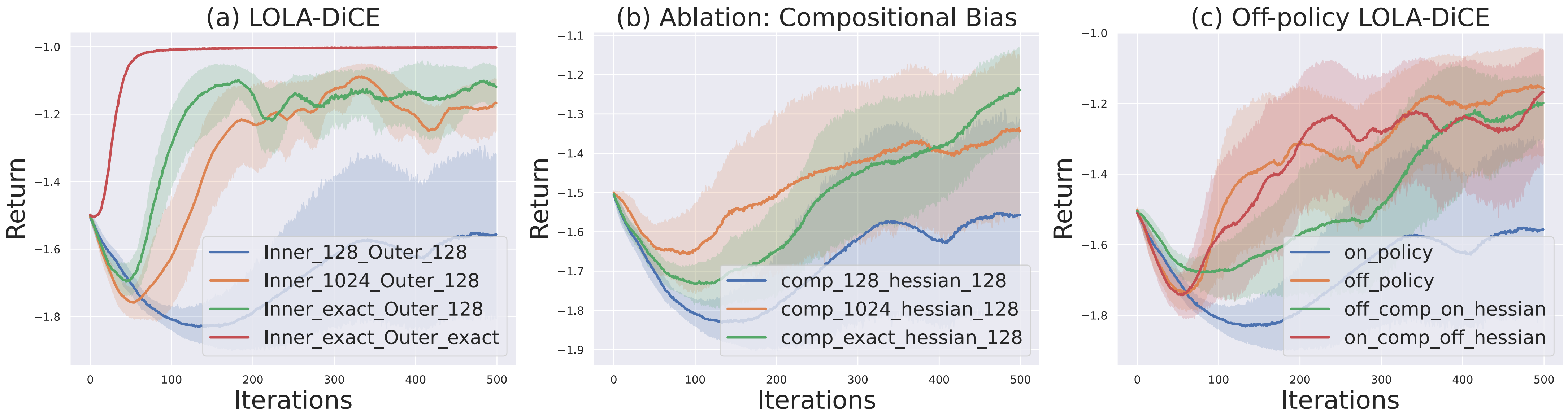


Ablation study of meta-gradient bias due to the compositional bias in different estimators, step sizes, learning rate

Ablation study of meta-gradient bias due to the Hessian bias in different learning rates and Hessian bias coefficients

Empirical Results

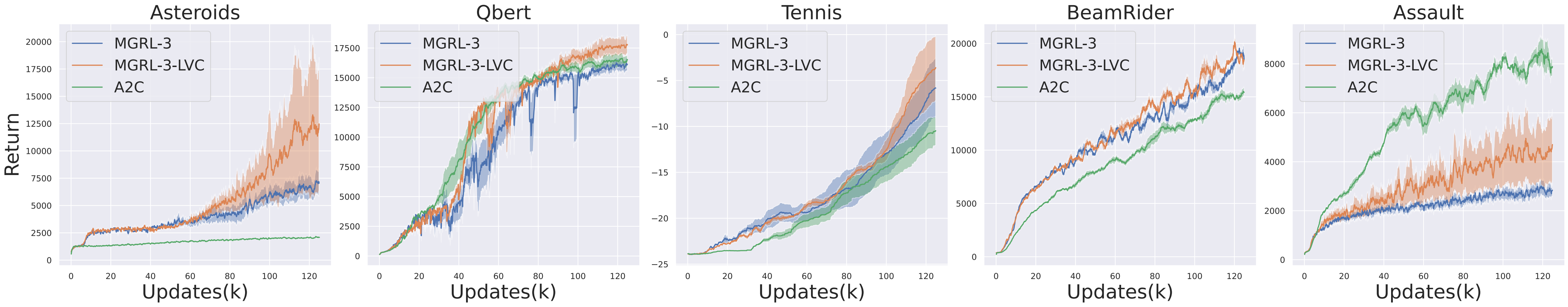
- Iterated Prisoner Dilemma (IPD)



Experiment result of LOLA-DiCE over 10 seeds

Empirical Results

- Atari 2600



Experimental results on Atari games over 5 random seeds.

Code Reference

- TorchOpt: https://github.com/metaopt/torchopt
- OpTree:https://github.com/metaopt/optree
- https://github.com/alexis-jacq/LOLA_DiCE

- Code Release:
- https://github.com/Benjamin-eecs/Theoretical-GMRL

Thank you!

benjaminliu.eecs@gmail.com
xidong.feng.20@ucl.ac.uk
yaodong.yang@pku.edu.cn